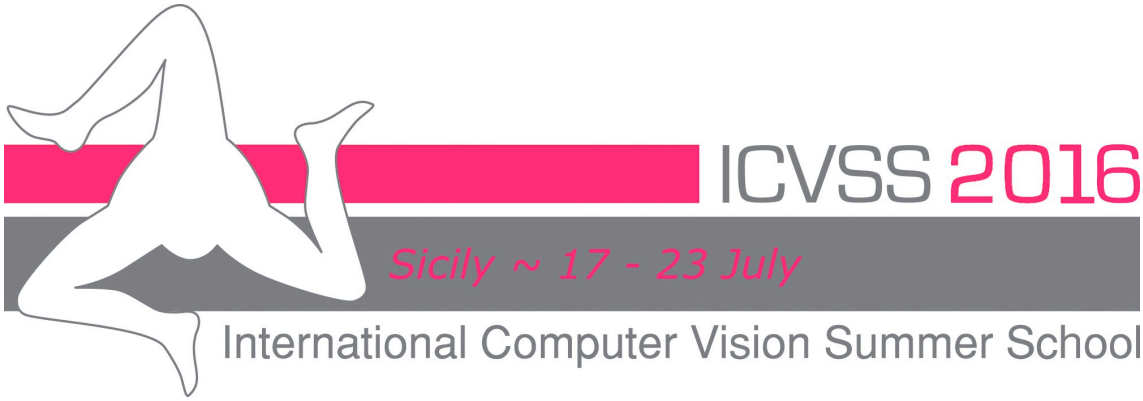

Towards Transparent AI Systems

Dhruv Batra
Virginia Tech, USA

Abstract

When today's intelligent systems fail, they typically fail in a spectacularly disgraceful manner, without warning or explanation, leaving the user staring at an incoherent output, wondering why the system did what it did. The root cause is lack of transparency. With a few rare exceptions, the emphasis in machine learning and AI communities today is on building systems with good predictive performance, not transparency. As a result, the users of these intelligent systems perceive them as inscrutable black boxes that cannot be understood or trusted.

In this talk, I will discuss research taking steps towards answering the difficult but important question of — "Why does an intelligent system do what it does?". I will use applications such as image classification, image captioning, visual question answering as test-beds and show techniques for creating visualizations, attention models, and human studies to compare machines and humans.

Keywords
Transparency, Interpretability, Deep Learning, VQA